On minimizing distortion and relative entropy

Michael P. Friedlander and Maya R. Gupta

December 2003

***Abstract*— A common solution for estimating a probability mass function w when given a prior pmf q and a linear constraint $Aw = b$ is to minimize the relative entropy between w and q subject to the linear constraint. In such cases, the solution w is known to have exponential form. We consider the case that the linear constraint is noisy, uncertain, infeasible, or otherwise "soft." A solution can be obtained by minimizing both the relative entropy and the distortion between Aw and b. A penalty parameter σ weights the relative importance between these two objectives. We show that this penalty formulation also yields a solution w with exponential form. If the distortion is based on an ℓ_p norm, then the exponential form of w is shown to have exponential decay parameters that are bounded as a function of σ. We also state conditions under which the solution w to the penalty formulation will result in zero distortion, so that the linear constraint $Aw = b$ holds exactly. These properties are useful in choosing penalty parameters, evaluating the impact of chosen penalty parameters, and proving properties about methods that use such penalty formulations.**

***Index Terms*— relative entropy, Kullback-Leibler distance, maximum entropy, penalty formulation, inverse problem, convex optimization, exact penalty function, moment constraint**

I. Introduction

CONSIDER the problem of estimating a probability mass function (pmf) $w \in \mathbb{R}^k$ given a strictly positive prior $q \in \mathbb{R}^k$. A common restriction on w is a mean constraint, so that if there are k observations $A = [a_1, \ldots, a_k] \in \mathbb{R}^{d\times k}$ and a mean $b \in \mathbb{R}^d$, then w must satisfy $Aw = b$. A standard approach [1], [2], [3] is to minimize the relative entropy function

$$\mathcal{I}(w; q) = \sum_{j=1}^{k} w_j \log \frac{w_j}{q_j} \tag{1}$$

over the constrained probability simplex

$$\mathbf{1}^T w = 1, \qquad w \geq 0, \tag{2}$$

$$Aw = b, \tag{3}$$

where the symbol $\mathbf{1}$ denotes a vector of ones; its length is determined by context. Often the prior q is the uniform distribution (i.e., $q_i = 1/k$ for each $i = 1, \ldots, k$); the minimization of (1) is then equivalent to maximizing entropy [4], [5], [6].

Preprint ANL/MCS-P1110-1203, December 2003, Mathematics and Computer Science Division, Argonne National Laboratory

M. P. Friedlander is with the Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439-4844 (e-mail: michael@mcs.anl.gov). This author's work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy contract W-31-109-Eng-38.

M. R. Gupta is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 (e-mail: gupta@ee.washington.edu). This author's work was supported in part by the National Science Foundation under Grant Number CCR-0073050

We recognize that the data may be noisy or uncertain; or the constraints (2)–(3) may be *infeasible* and admit no solution. A more appropriate estimate of w may then be the solution to

$$\begin{array}{ll} \underset{w}{\text{minimize}} & \mathcal{I}(w; q) + \sigma D(Aw - b) \\ \text{subject to} & \mathbf{1}^T w = 1, \qquad w \geq 0, \end{array} \tag{4}$$

where $D : \mathbb{R}^d \to \mathbb{R}$ is a convex function that measures the distortion in satisfying (3), and σ is a positive scalar that can be used to balance the tradeoff between minimizing $\mathcal{I}(w; q)$ and $D(Aw - b)$. The parameter σ can be set in response to expected noise in measurements of A or b and may be set differently depending on the units of measurement of the distortion function compared with the bits of relative entropy.

In this paper we characterize the analytic form of the minimizer of (4). We show that for any convex distortion D, the solution has an exponential form. Moreover, for a family of distortion functions based on the ℓ_p norm, the parameters of the exponential solution can be further characterized: the rate of decay is bounded by a function of σ. These results extend the work of Campbell [7], who assumed a scalar feature space (i.e., $d = 1$), and D as the ℓ_1 distortion on the mean constraint, so that $D(Aw - b) = \|Aw - b\|_1$.

The exponential form of the solutions to constrained minimum relative entropy problems has been known for many years. Proof of an exponential form for a continuous problem and general constraints can be found in Kullback [2, Theorem 2.1] and Cover and Thomas [4, Chapter 11]. Similar problems arise in rate distortion theory, and solutions in that framework have also been shown to have an exponential form. (For an overview of rate-distortion results, see [4] and [8].)

The results in this paper grant deeper insight into the estimates resulting from problems of the form (4). In particular, we see how the choice of σ directly affects the estimate. In Section IV we show that for a family of distortion functions D based on the ℓ_p norm, a minimizer to (4) will satisfy the linear constraint exactly for all σ over a finite threshold value.

The bounded exponential decay property shows that the estimator will weight the observations a_i (the columns of A) so that no observation receives an arbitrarily large or small relative weight, regardless of the number of observations or their specific values. Gupta et al. [9] use this bounded exponential decay property to prove the statistical consistency of distributions estimated with the linear interpolation with maximum entropy (LIME) nonparametric statistical learning algorithm. The LIME algorithm is a neighborhood method that weights each neighborhood training sample with an adaptive kernel that solves a problem of the form (4) with a uniform prior q. Our results presented here show that the form of the LIME adaptive kernel must be exponential. The bounded exponential decay given in Theorem 4.2 translates into a bound

on the ratio between the largest and the smallest components of the LIME weight kernel. The bounded ratio allows one to state robust properties about the LIME weight kernel's behavior even though the neighborhood training samples (the columns of A) are random. In particular, the bounded ratio is important in order to verify Stone's [10] conditions for the consistency of any neighborhood learning method.

Inference methods based on minimum relative (or maximum) entropy with constraints are now classic [11], [2], [12], [1] and enjoy continuing popularity [13], [3], [14], [15], [16]. There has also been recent interest in approaches that do not necessarily treat (3) as a *hard* constraint (see, e.g., [7], [17], [18]).

The optimization problem (4) is *convex*—both $\mathcal{I}$ and D are convex, and the constraints are linear. If D is continously differentiable, its convex structure makes it especially suitable for numerical solution by a variety of interior-point solvers for nonlinear optimization such as KNITRO [19], LOQO [20], and MOSEK [21]. Although in practice a numerical solution can be computed, theoretical understanding of an estimation method based on (4) relies on a description of how the solution depends on the problem parameters.

Previously, Campbell [7] gave an analytic solution for the minimizing pmf of (4) for the scalar case—when A is a $1 \times k$ matrix and b is a scalar—for the ℓ_1 distortion under the assumption that the uncertain mean b lies within the range of the set of events $\min_i a_i \leq b \leq \max_i a_i$. We extend Campbell's result in two useful ways. First, the data vectors can be of arbitrary finite dimension instead of one dimensional, so that instead A has d rows. Second, the mean b need not be in the convex hull of the data vectors specified by the columns of A, so that the solution still exists even if the constraints (2)–(3) are infeasible.

II. The exponential form

Before we consider the penalty function formulation of the minimum relative entropy problem, we first examine the case where the constraints (2)–(3) are imposed explicitly:

$$\begin{array}{ll} \underset{w}{\text{minimize}} & \mathcal{I}(w;q) \\ \text{subject to} & \mathbf{1}^T w = 1 \\ & Aw = b. \end{array} \tag{5}$$

We have anticipated a strictly positive solution w^* and disregarded the nonnegativity constraint $w \geq 0$, though in practice they would be explicit.

In this section we establish that the solution of (5) is exponential; its parameters are the columns of A and the Lagrange multipliers associated with the second constraint. (The Lagrange multipliers of the constraint $\mathbf{1}^T w = 1$ can be eliminated.) As discussed in the introduction, this result is not new, and it can be derived from a variety of perspectives. However, it is useful to show a complete derivation so that we may refer to it later. One elegant approach to derive this result is to consider the dual of (5) (every convex optimization problem has a dual). The recent book by Boyd [22] provides an accessible discussion on this topic.

The optimization problem (5) is *convex*; in other words its objective function is convex, and the equality constraints are linear. Under a suitable *constraint qualification*, such as Slater's condition, the first-order optimality conditions of (5) are in fact both necessary and sufficient (see, for example, [23], [22]).

Definition 2.1 (Slater's Condition): There exists a point w in the relative interior of the feasible set.

Applied to (5) (and recalling the implicit nonnegativity constraint on w), Slater's condition implies that there exists a w that satisfies

$$w > 0, \qquad \mathbf{1}^T w = 1, \qquad Aw = b.$$

Let ζ and y be the Lagrange multipliers associated with the first and second constraints of (5), respectively. An optimal point, together with its associated Lagrange multipliers, must satisfy the Karush-Kuhn-Tucker (KKT) conditions.

Definition 2.2 (First-Order KKT Optimality Conditions): A triple (w^*, ζ^*, y^*) is a first-order KKT point of (5) if it satisfies the following conditions:

$$\mathbf{1}^T w = 1 \tag{6a}$$

$$Aw = b \tag{6b}$$

$$\nabla_w \mathcal{I}(w,q) + \zeta \mathbf{1} + A^T y = 0. \tag{6c}$$

Theorem 2.3: Suppose that Slater's condition holds. Then the vector with components

$$w_j^* = \frac{u_j}{\sum_{j=1}^k u_j}, \tag{7}$$

for $j = 1, \ldots, k$, solves (5), where

$$u_j = q_j \exp(-a_j^T y^*), \tag{8}$$

and y^* is the Lagrange multiplier corresponding to the constraint $Aw = b$.

Proof: By Slater's condition, the feasible set of (5) is nonempty. Moreover, the level sets $S_{w_0} = \{w \mid \mathcal{I}(w;q) \leq \mathcal{I}(w_0;q)\}$ are closed and bounded. The strict convexity of $\mathcal{I}$ therefore implies that there exists a unique solution w^* to (5). Slater's condition is sufficient to guarantee that there exist Lagrange multipliers ζ^* and y^* such that (w^*, ζ^*, y^*) satisfies the KKT conditions (6).

Note that $\nabla_w \mathcal{I}(w;q)_j = 1 + \log(w_j/q_j)$. Solve the jth equation of (6c) for w_j^* to obtain

$$w_j^* = q_j \exp(-\zeta^* - a_j^T y^* - 1). \tag{9}$$

Sum (9) over all j, and use (6a) to obtain

$$\sum_{j=1}^k q_j \exp(-\zeta^* - a_j^T y^* - 1) = 1.$$

Hence, ζ^* must satisfy

$$\zeta^* = \log \left(\sum_{j=1}^k q_j \exp(-a_j^T y^* - 1) \right). \tag{10}$$

Replacing ζ^* in (6c) with (10), and subsequently solving for w_j^*, we arrive at

$$w_j^* = \frac{q_j \exp(-a_j^T y^*)}{\sum_{j=1}^k q_j \exp(-a_j^T y^*)}, \tag{11}$$

as required. ■

III. Penalty formulation

It may be that the constraints (2)–(3) are infeasible, or that the constraint $Aw = b$ need not be solved exactly. For example, the data may be known to be noisy such that $Aw = b + n$, where n is some known or unknown noise; or the mean constraint may be uncertain; or fidelity to the prior q may be highly important relative to the constraint. These cases may be captured by introducing the set of constraints $Aw = b$ into the objective via a penalty function as done in (4).

In Lemma 3.1 we show that the solution to (4) will have an exponential form if D is convex. The simple proof of this lemma provides an insight: the weights that solve (4) are the same as the weights that solve (5) with the mean constraint $Aw = \bar{b}$ for some $\bar{b}$.

Lemma 3.1: Suppose D convex. Then there exists a unique solution w^* to (4) with components that satisfy (7) and (8) where y^* is the Lagrange multiplier corresponding to the solution of (5) with the constraint $Aw = \bar{b}$, for some $\bar{b}$.

Proof: Over the compact set defined by the constraints $\mathbf{1}^T w = 1$ and $w \geq 0$, D is convex (and therefore continuous) and I is strictly convex and continuous. Therefore, a unique minimizer w^* of (4) exists.

Suppose the unique minimizer w^* of (4) is known. Let $\bar{b} = Aw^*$. Solve (5) with the constraint $Aw = \bar{b}$ for a minimizer $w^\sharp$ (such a minimizer must exist because the constraint is feasible for w^*). Since $w^\sharp$ solves (5), it satisfies the constraint $Aw^\sharp = \bar{b}$, and thus $D(Aw^\sharp - \bar{b}) = 0 = D(Aw^* - \bar{b})$.

Further, it must be that $\mathcal{I}(w^*; q) = \mathcal{I}(w^\sharp; q)$, because if otherwise $\mathcal{I}(w^*; q) < \mathcal{I}(w^\sharp; q)$ then $w^\sharp$ could not solve (5) because w^* would satisfy the constraint $Aw^* = \bar{b}$ and have lower relative entropy. Similarly, assuming $\mathcal{I}(w^*; q) > \mathcal{I}(w^\sharp; q)$ requires that w^* not be the minimizer of (4).

Then since w^* satsifies the constraint $Aw^* = \bar{b}$ and $\mathcal{I}(w^*; q) = \mathcal{I}(w^\sharp; q)$, it must be a minimizer of (5). Since $w^\sharp$ is defined as the minimizer of (5), and there is a unique minimizer, then w_j^* must equal $w_j^\sharp$ for all j.

From Theorem 2.3, $w^\sharp$ has the form defined by (7) and (8), and thus w^* must also have the form defined by (7) and (8). ■

IV. Exact penalty formulation

When the distortion function takes the form

$$D(Aw - b) = \|Aw - b\|_p, \tag{12}$$

with $1 \leq p \leq \infty$, the penalty function formulation (4) is *exact*—for a finitely large penalty parameter σ, its solution has zero distortion, so that the constraint (3) is satisfied exactly. With an exact penalty function we can construct a function whose unconstrained optima coincide with the optima of the constrained problem. Exact penalty functions play an important role in the modeling of continuous optimization problems; they have rich theoretical properties, and when the norm is polyhedral (i.e., $p = 1$ or $p = \infty$), they are computationally practical because they can be reformulated as polyhedral constraints. They share the same philosophy with the more general class of penalty functions: by augmenting the objective function to include a penalty on the constraint violation, a constrained (and possibly difficult) problem can be transformed into an unconstrained (and easier, we hope) problem. Exact penalty functions were first analyzed by Pietrzykowski [24], and later by Bertsekas [25], Fletcher [26], and Han and Mangasarian [27], among others.

For the remainder of this section, let the penalty function D in (4) be of the form given by (12). Note that the objective of (4) is convex for any $1 \leq p \leq \infty$; but it is not everywhere differentiable (i.e., it is only $\mathbb{C}^0$). When the norm is polyhedral, (4) can be reformulated as an equivalent and smooth problem, and in that case, the corresponding first-order KKT conditions (see Definition 2.2) can be applied to find optimal solutions. Moreover, a variety of algorithms for smooth, constrained optimization could then be used to numerically solve the smooth reformulations. We discuss one such reformulation for $p = 1$ in Section IV-C. However, there exists a rich theory of optimization for nonsmooth functions, and a result analogous to Theorem 2.3 can be derived for (4) (see Theorem 4.2).

A. Nonsmooth optimality concepts

We summarize in this section some of the optimality concepts from nonsmooth optimality that we need for our analysis. Our treatment follows the approach of [26]. The vector g is a *subgradient* of the convex function $f(x) : \mathbb{R}^n \to \mathbb{R}$ at x if

$$f(x + p) \geq f(x) + g^T p,$$

for all $p \in \mathbb{R}^n$. The subgradient is normal to a supporting hyperplane of f at x. The set of all subgradients

$$\partial f(x) \stackrel{\text{def}}{=} \{g \mid f(x + p) \geq f(x) + g^T p \quad \text{for all} \quad p \in \mathbb{R}^n\}$$

is the *subdifferential* of f at x. When f is differentiable at x, there is only a single supporting hyperplane at that point, and the subgradient is unique and corresponds to $\nabla f(x)$.

Definition 4.1 (Nonsmooth First-Order Optimality): A triple (w^*, ζ^*, y^*) is a first-order optimal point of (4) if it satisfies the following conditions:

$$\mathbf{1}^T w = 1 \tag{13a}$$

$$\nabla_w \mathcal{I}(w; q) + \zeta \mathbf{1} + A^T y = 0, \tag{13b}$$

where $y^* \in \partial(\sigma \|Aw^* - b\|_p)$.

Comparing (6) with (13), we see how $y^* \in \partial(\sigma \|Aw^* - b\|_p)$ may be interpreted as a kind of Lagrange multiplier for the constraint $Aw = b$ implied by the penalty function (12). Note that for any given σ, $Aw^* - b$ may or may not equal zero.

The *dual norm* is particularly important for the analysis of exact penalty functions and will be useful in our analysis. For

any norm $\|\cdot\|$ in $\mathbb{R}^n$, the corresponding dual norm is defined as

$$\|y\|_D = \sup_{\|x\|\leq 1} y^T x.$$

For any p and q such that $1/p+1/q = 1$, the ℓ_p and ℓ_q norms are dual to each other.

B. Solution of the exact penalty function

Theorem 4.2 shows that penalty function formulation (4) always has a solution with the form specified in (7) and (8), and exponenital parameter y^* which is bounded as a function of σ.

Theorem 4.2: There exists a vector w^*, with components defined by (7) and (8), that solves (4). The parameter $y^* \in \partial(\sigma\|Aw^* - b\|_p)$.

Proof: The feasible set of (4) is nonempty, and the level sets $S_{w_0} = \{w \mid \mathcal{I}(w;q) \leq \mathcal{I}(w_0;q)\}$ are closed and bounded. Hence, the strict convexity of $\mathcal{I}$ implies that there exists a unique solution w^* to (4). Moreover, the constraints $\mathbf{1}^T w = 1$ are linear so that, by [26, Theorem 14.6.1], there exist a vector $y^* \in \partial(\sigma\|Aw^* - b\|_p)$ and a scalar ζ^* such that (w^*, ζ^*, y^*) satisfies the first-order optimality conditions (13).

The form of the solution w^* can be derived in the same manner as (11). ∎

Note that the condition $y^* \in \partial(\sigma\|Aw^* - b\|_p)$ imposes an implicit bound on the magnitude of y^*. With the definition of the dual norm, the subdifferential of $\sigma\|Aw - b\|_p$ can be equivalently stated (see [26, Chapter 14]) as

$$\partial(\sigma\|Aw-b\|_p) = \{y \mid y^T(Aw-b) = \sigma\|Aw-b\|_p, \|y\|_D \leq \sigma\}, \tag{14}$$

and so σ provides a bound on the dual norm of y^*. This creates a bound on the exponential decay of the solution w^* by virtue of (11).

The exact penalty function formulation is exact in the following sense: for all values of the penalty parameter over a certain threshold value, KKT points of (5) are also stationary points of its exact penalty function formulation.

Theorem 4.3: Let w^* be a solution of (5), with corresponding Lagrange multipliers ζ^* and y^* (see Theorem 2.3). Then for every $\sigma > \|y^*\|_D$, w^* is also a minimizer of (4).

Proof: This result follows immediately from Theorem 14.3.1 in [26]. ∎

C. The ℓ_1-penalty function

When the penalty function of (4) uses the ℓ_1 norm, its objective is discontinuous over sets of hyperplanes. However, there is a common device in the optimization literature for reformulating it as an equivalent and smooth problem (see, for example, [27, Theorem 4.8] and [28, Section 4.2.3]). We introduce a set *elastic variables* $r, s \geq 0$, and rewrite (5) as

$$\begin{array}{ll} \underset{w,r,s}{\text{minimize}} & \mathcal{I}(w;q) + \sigma\mathbf{1}^T(r+s) \\ \text{subject to} & \mathbf{1}^T w = 1 \\ & Aw + r - s = b \\ & r, s \geq 0. \end{array} \tag{15}$$

The solution of (15) is a 5-tuple $(w^*, r^*, s^*, \zeta^*, y^*)$ that satisfies the first-order KKT conditions

$$\mathbf{1}^T w = 1 \tag{16a}$$
$$Aw + r - s = b \tag{16b}$$
$$\nabla_w \mathcal{I}(w;q) + \zeta\mathbf{1} + A^T y = 0 \tag{16c}$$
$$\min(r, \sigma\mathbf{1} - y) = 0 \tag{16d}$$
$$\min(s, \sigma\mathbf{1} + y) = 0. \tag{16e}$$

The last two conditions (16d)–(16e) imply that their arguments are nonnegative, so that $\sigma\mathbf{1} \geq y \geq -\sigma\mathbf{1}$. This pair of inequalities can be conveniently restated as

$$\|y\|_\infty \leq \sigma. \tag{17}$$

Note that the ℓ_1 and ℓ_∞ norms are dual to each other, so that (17) is equivalent to $\|y\|_D \leq \sigma$ (see (14)).

Lemma 4.4: Suppose that $(w^*, r^*, s^*, \zeta^*, y^*)$ is a solution of (16) with $\sigma > 0$. Then r^* and s^* are componentwise complementary; that is, $r_i^* s_i^* = 0$, for $i = 1, \ldots, d$.

Proof: Set $z^r = \sigma\mathbf{1} - y^*$ and $z^s = \sigma\mathbf{1} + y^*$. Then

$$z^r + z^s = 2\sigma\mathbf{1} > 0 \tag{18}$$

because $\sigma > 0$ by hypothesis. Note that (16d)–(16e) imply that

$$r_i^* z_i^r = s_i^* z_i^s = 0. \tag{19}$$

Now suppose that $r_i^* > 0$. Multiplying the ith component of (18) by r_i^* yields $r_i^* z_i^s > 0$. Hence $z_i^s > 0$, and from (19), we have $s_i^* = 0$. By analogous argument, $s_i^* > 0$ implies $r_i^* = 0$. Then $r_i^* s_i^* = 0$, as required. ∎

Theorem 4.5 (Exponential form): Let σ be a positive constant. Suppose that $(w^*, r^*, s^*, \zeta^*, y^*)$ is a solution of (15). Let $\mathcal{A}, \mathcal{A}_+, \mathcal{A}_-$ be index sets such that

$$\begin{aligned} (Aw)_i &= b, \quad i \in \mathcal{A} \\ (Aw)_i &> b, \quad i \in \mathcal{A}_+ \\ (Aw)_i &< b, \quad i \in \mathcal{A}_-. \end{aligned}$$

Then

$$w_j^* = \frac{u_j}{\sum_{j=1}^k u_j},$$

where

$$u_j = q_j \exp\left(\sum_{i\in\mathcal{A}} a_{ij} y_j^* + \sigma \sum_{i\in\mathcal{A}_+} a_{ij} - \sigma \sum_{i\in\mathcal{A}_-} a_{ij}\right). \tag{20}$$

Proof: Because (15) and (4) are equivalent, w^* must have the form specified by (7)–(8).

Now we consider the values that each y_i^* may have. If $i \in \mathcal{A}_+$, then $(Aw^*)_i > b$, and (16b) together with (16d) implies that $0 \leq r_i^* < s_i^*$. By Lemma 4.4, we must have $r_i^* = 0$. Then from (16e) we deduce that $y_i^* = -\sigma$. By analogous argument, $y_i^* = \sigma$ for $i \in \mathcal{A}_-$. For $i \in \mathcal{A}$, $(Aw^*)_i = b$, and by (16b), $r_i^* = s_i^*$. Lemma 4.4 then implies $r_i^* = s_i^* = 0$, and so we have from (16d)–(16e) that $\sigma \geq y_i^* \geq -\sigma$. In summary, we may now write $a_j^T y^*$ as

$$a_j^T y^* = \sum_{i\in\mathcal{A}} a_{ij} y_j^* + \sigma \sum_{i\in\mathcal{A}_+} a_{ij} - \sigma \sum_{i\in\mathcal{A}_-} a_{ij}, \tag{21}$$

for each $j = 1, \ldots, k$. Substituting (21) into (8), we see that w_j^* has the required form. ∎

ACKNOWLEDGMENTS

We would like to thank L. Lorne Campbell, Robert M. Gray, and Richard A. Olshen for helpful discussions. In particular, L. Lorne Campbell's careful reading of this paper helped clarify the description and led to many detailed changes.

REFERENCES

[1] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. 33, no. 1, pp. 26–37, 1980.

[2] S. Kullback, *Information Theory and Statistics*. New York: John Wiley and Sons, 1959.

[3] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.

[4] T. Cover and J. Thomas, *Elements of Information Theory*. United States of America: John Wiley and Sons, 1991.

[5] N. Wu, *The Maximum Entropy Method*. Berlin: Springer-Verlag, 1997.

[6] G. Erickson and C. R. Smith, *Maximum Entropy and Bayesian Methds in Science and Engineering*. U.S.A.: Kluwer Academic Publishers, 1988.

[7] L. L. Campbell, "Minimum cross-entropy estimation with inaccurate side information," *IEEE Trans. Inform. Theory*, vol. 45, no. 7, pp. 2650–2652, November 1999.

[8] R. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley and Sons, 1968.

[9] M. Gupta, R. M. Gray, and R. Olshen, "Nonparametric supervised learning with reduced bias," Submitted for possible publication. Preprint available at www.ee.washington.edu/people/faculty/gupta_maya/GGO03.pdf, 2003.

[10] C. Stone, "Consistent nonparametric regression," *The Annals of Statistics*, vol. 5, no. 4, pp. 595–645, 1977.

[11] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, pp. 620–630, 1957.

[12] J. P. Burg, "Maximum entropy spectral analysis," *37th Annual International Meeting of the Society of Exploratory Geophysics*, 1967.

[13] J. Navaza, "The use of non-local constraints in maximum-entropy electron density reconstruction," *Acta Crystallographica*, pp. 212–223, 1986.

[14] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," *Advances in Neural Information Processing Systems 12*, 1999.

[15] J.-F. Bercher, G. LeBesnerais, and G. Demoment, "The maximum entropy on the mean method, noise, and sensitivity," *Maximum Entropy and Bayesian Methods*, pp. 223–232, 1996.

[16] H. Gzyl and Y. Velasquez, "Maxentropic interpolation by cubic splines with possibly noisy data," *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 20th International Workshop*, pp. 216–228, 2001.

[17] G. L. Besenerais, J.-F. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1565–1577, 1999.

[18] I. Csiszár, F. Gamboa, and E. Gassiat, "MEM pixel correlated solution for generalized moment and interpolation problems," *IEEE Trans. Inform. Theory*, vol. 45, no. 7, pp. 2253–2270, November 1999.

[19] R. Byrd, M. E. Hribar, and J. Nocedal, "An interior point method for large scale nonlinear programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 877–900, 1999.

[20] D. Shanno and R. Vanderbei, "An interior-point algorithm for nonconvex nonlinear programming," *Computational Optimization and Applications*, vol. 13, pp. 231–252, 1999.

[21] E. D. Andersen, "Mosek: `http://www.mosek.com/documentation.html`," 2003.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

[23] O. L. Mangasarian, *Nonlinear Programming*, ser. Classics in Applied Mathematics, G. Golub, Ed. Philadelphia: SIAM, 1994, vol. 10, originally published: New York, McGraw-Hill, 1969.

[24] T. Pietrzykowski, "An exact potential method for constrained maxima," *SIAM J. Numer. Anal.*, vol. 6, pp. 262–304, 1969.

[25] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic Press, 1982.

[26] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: John Wiley and Sons, 1987.

[27] S.-P. Han and O. L. Mangasarian, "Exact penalty functions in nonlinear programming," *Math. Prog*, vol. 17, pp. 251–269, 1979.

[28] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. San Diego, California: Academic Press, 1981.